Diagnosis of Performance Degradation in Leadership Class (PetaScale) Systems

Kevin Harms – Argonne Leadership Computing Facility

Quick Facts

- Software Developer who works on the I/O subsystem
  - Part of LCF Operations Team
- ALCF Blue Gene P is called Intrepid
- Contains 40 BG/P racks
- Delivers approximately 557 TeraFLOPS
- Number 9 on the Top 500 (June 2010)
- Most the data presented here was gathered while working on performance tuning of our large parallel file system using PVFS

Photo

Motivation

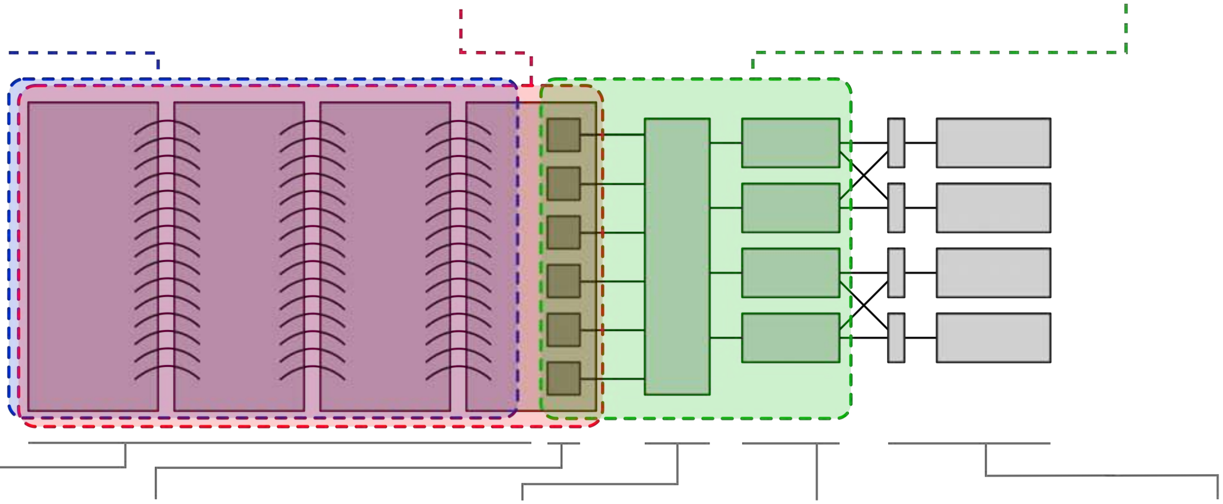
- Performance is important in order to meet science goals
- ALCF runs a "capability" machine
  - Fewer jobs that use more of the machine
- I/O is overhead for the most part
  - Checkpoints created in order to resume in case of failure
  - Smaller data for analysis
  - Time spent performing I/O is time spent not computing
- I/O synchronizes at some point
  - Collective I/O – MPI_File_write_all
  - Barrier waiting for all processes to complete individual I/O
  - I/O system becomes as fast as the slowest part

Intrepid Hardware Architecture

High-level I/O libraries and **MPI-IO** execute on compute nodes and organize accesses before the I/O system sees them.

I/O forwarding software runs on compute and I/O nodes and bridges between the compute nodes and external storage.

PVFS code runs on I/O and storage nodes, maintains logical storage space, and enables efficient access to data.



Compute nodes

40,960 Quad core PowerPC 450 nodes with 2 Gbytes of RAM each

I/O nodes

640 Quad core PowerPC 450 nodes with 2 Gbytes of RAM each

Commodity network

900+ port 10 Gigabit Ethernet Myricom switch complex

Storage nodes

128 two dual core Opteron servers with 8 Gbytes of RAM each

Enterprise storage

16 DataDirect S2A9900 controller pairs with 480 1 Tbyte drives and 8 InfiniBand ports per pair

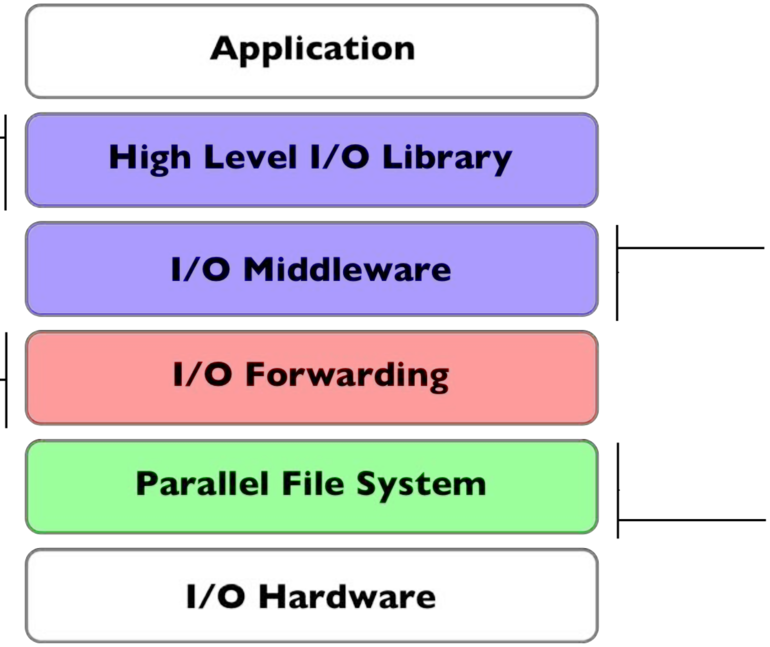
Intrepid Software Architecture

High Level I/O Library
maps application abstractions
onto storage abstractions and
provides data portability

HDF5, Parallel NetCDF

I/O Forwarding
Bridges between application
tasks and storage system

IBM ciod



Application

High Level I/O Library

I/O Middleware

I/O Forwarding

Parallel File System

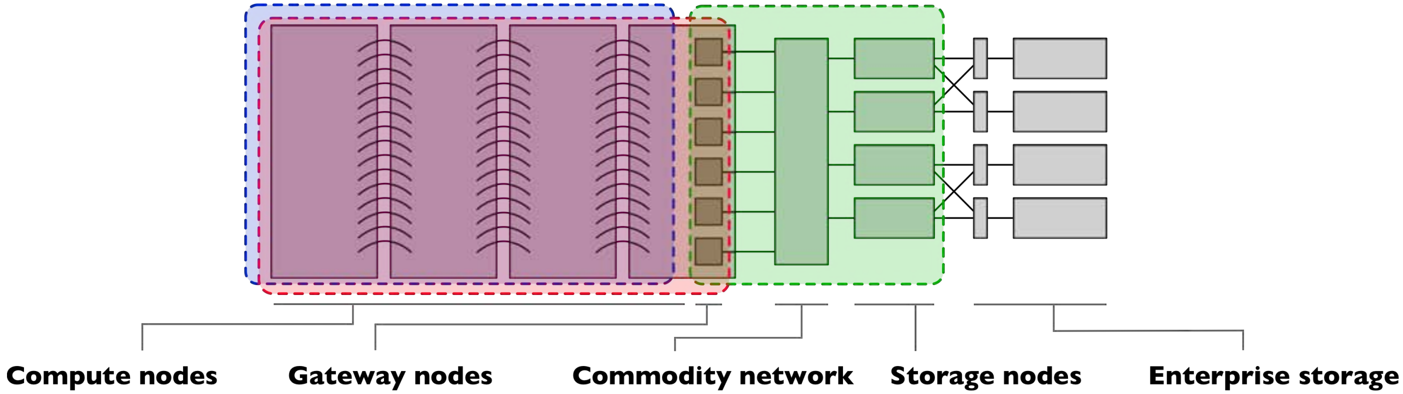
I/O Hardware

I/O Middleware
Organizes accesses from
many processes, especially
those using collective I/O

MPI-IO

Parallel File System
Maintains logical space and
provides efficient access to data

PVFS



BG/P RAS Monitoring

- BG/P has comprehensive sense points
- Data collected for environmentals, power, DRAM, cache, CPU, network
- Events generated are inserted in a large database with all other BG configuration data.
- Admins or users can get a list of events that were reported while the job was running specific to the hardware it was running on.
- Reasonably easy to isolate problems because a running job does not share resources with any other job. Hardware partitions are electrically isolated.

BG/P RAS Monitoring – DRAM Example

- User running a benchmark on Surveyor and on Intrepid
  - Identical binary used for both runs
  - No I/O done

```
surveyor: Elapsed time, pclks,s:     759712251596     893.7791195247
intrepid: Elapsed time, pclks,s:     373230946208     439.0952308329
```

- The same code ran twice as fast on the same hardware due to memory related errors. Correctness was not impacted.

BG/P RAS Monitoring - DRAM Example (2)

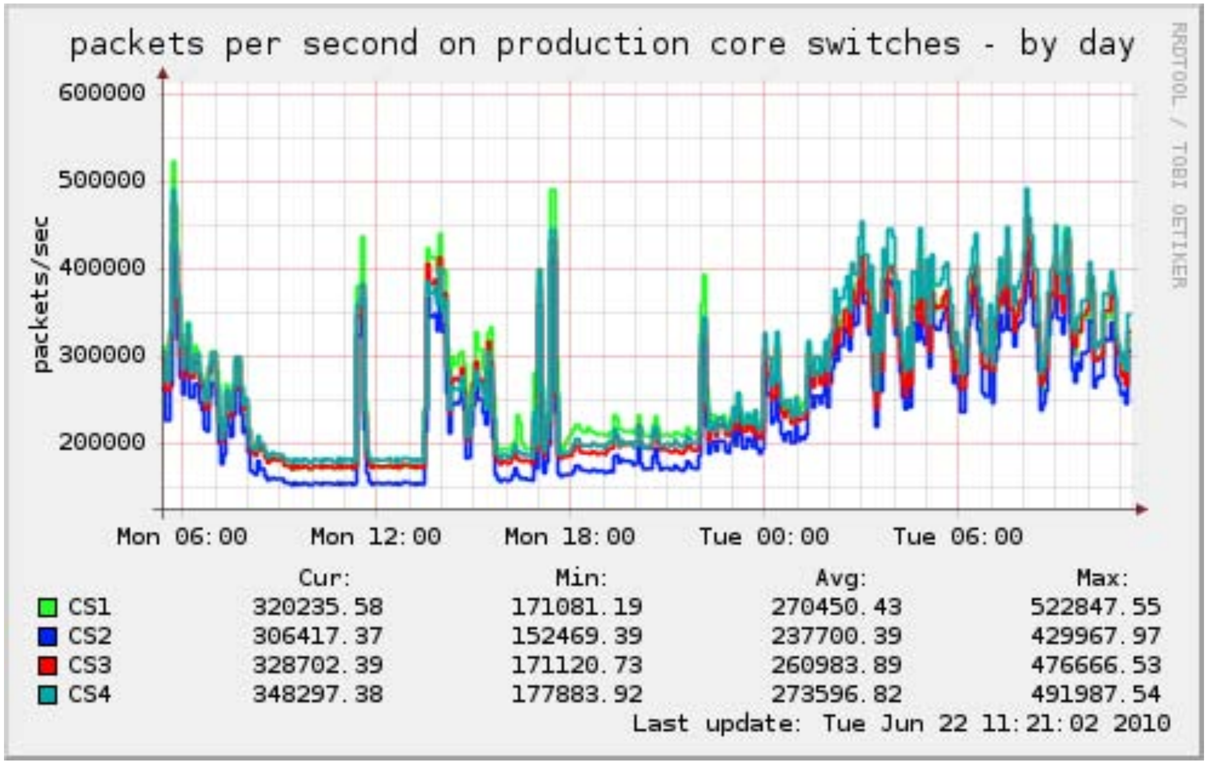
```
Time:            2010-05-11 14:58:27.601350
Record ID:       3879635
RAS Message ID:  KERN_0803
RAS Error Code:  _bgp_err_ddr_double_symbol_error
Block ID:        ANL-R00-M1-N08-256
Location:        R00-M1-N15-J20
BG Job ID:       287330
RAS Message Text:  DDR correctable double symbol error(s): DDR Controller 1, failing
     SDRAM address 0x03fd12a00, (1) BPC pin JG195, transfer 1, bit 17, BPC module pin
     AE24, compute trace MEMORY1DATA58, DRAM chip U19, DRAM pin D7.(2) BPC
     pin JL201, transfer 1, bit 23, BPC module pin F24, compute trace
     MEMORY0DATA57, DRAM chip U04, DRAM pin C2.
Severity:        WARN
```

The Numbers

- I/O Nodes – 640
- Myricom Switches – 10
- Myrinet Ports - 2416
- File Servers – 128
- Infiniband Ports – 128
- DDN 9900 Controllers – 32
- SATA Drives – 7680
- People assigned to I/O subsystem – 2

Myrinet Uplink Bandwidth

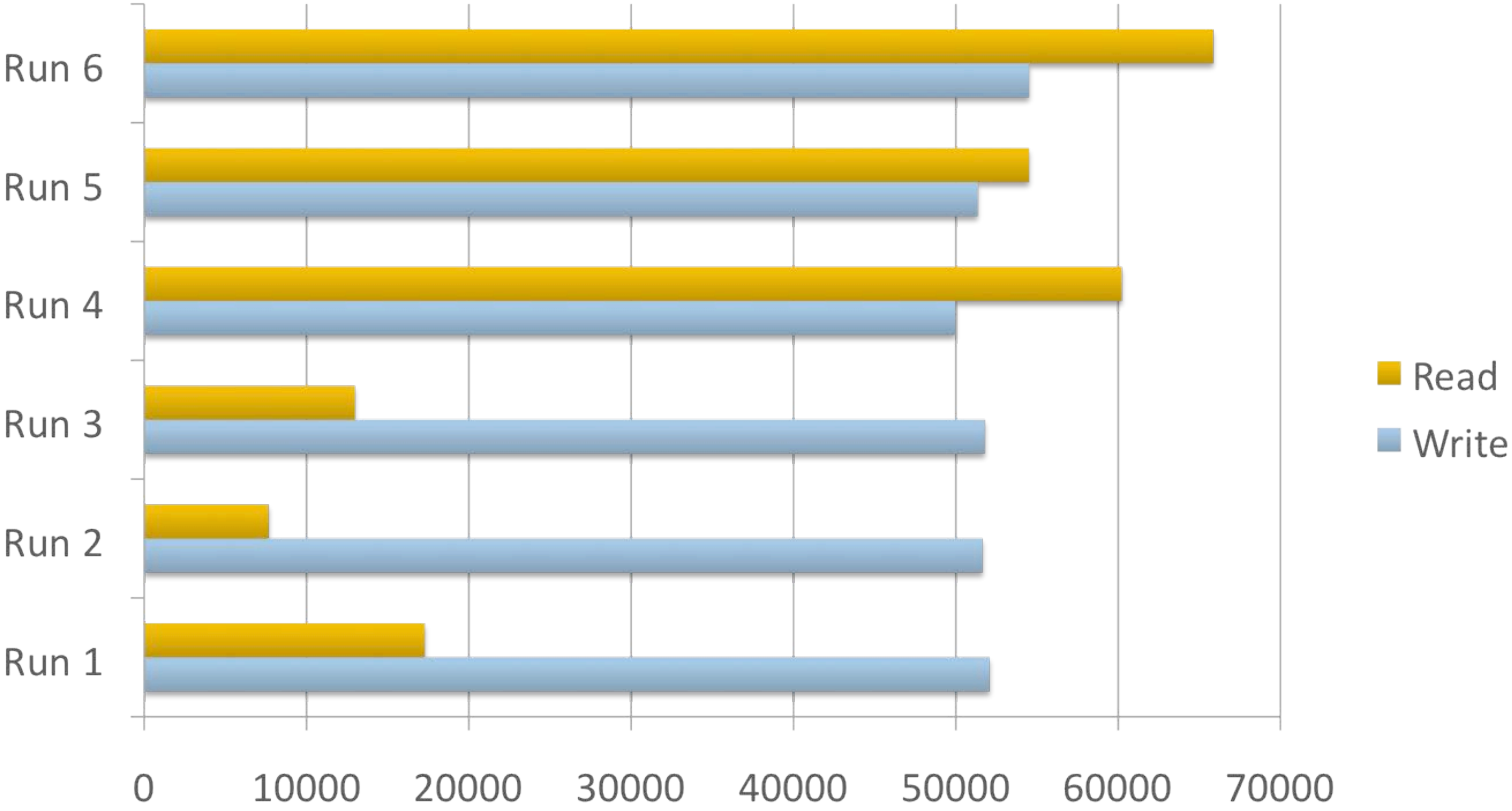
- Early in deployment there was an issue where edge switches would only use 1 core switch to uplink to instead of all 4.
- All operations would complete normally except significant loss in throughput.



Myrinet Bad CRC

- Failures of hardware can result in corruption of packets that will fail CRC checks.
  - Line cards and optics
- Myrinet will pass these bad frames along and they will be dropped at the destination
- Results in bad CRCs being reported at many different switch ports
- This type of fault could result in significant performance loss
- Difficult to track down which port is the source of the problem
- Solution: try disabling ports until the problem goes away

Myrinet Bad CRC (2)



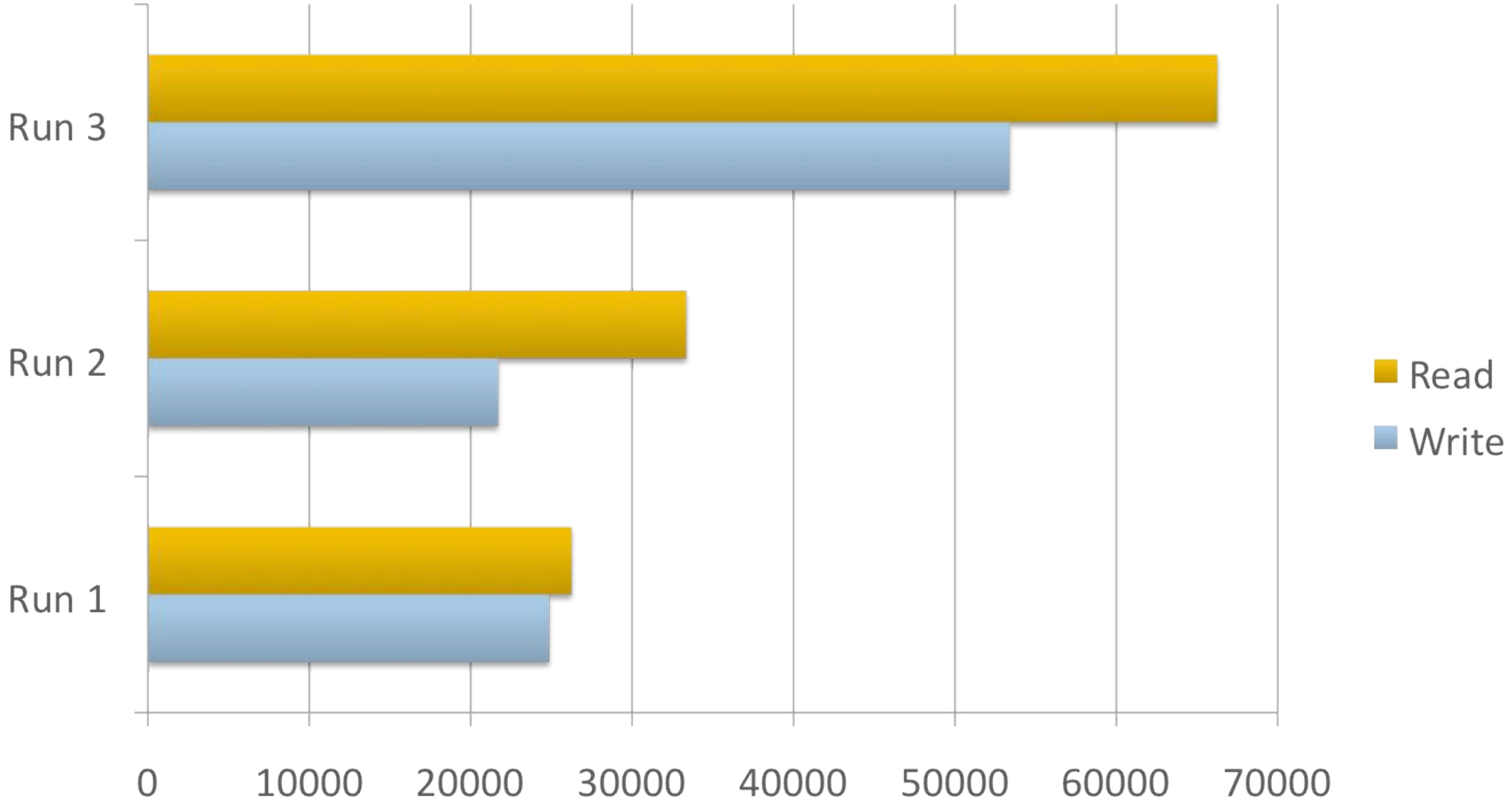
DDN Slow Disk Drive

- A single slow drive in one DDN can cause the entire parallel file system to become "slow" because of the aforementioned need to synchronize at some point.
- Intrepid uses SATA drives for cost reasons, however SATA drives have no performance guarantees associated with them.
- DDN provides a myriad of statistics about each disk, just need to examine 8000 disks.

DDN Slow Disk Drive (2)

```
                          Tier 1 Delay Statistics
 Time                           Disk Channels
seconds   A      B      C      D      E      F      G      H      P      S
  0.017  ba111  baa53  bb903  ba814  bad5a  ba452  bbf95  ba57c  bb06c  b9c9e
  0.033 166794 167079 166713 1662f7 165d73 16665b 167934 167ddd 166417 164df2
  0.050  866f0  85c2c  85e27  86599  867b4  8698e  857de  86873  861fb  86250
  0.067  30043  2f59b  2f60b  2fee4  2fd9d  2fdf9  2e896  2f112  2fa21  310e5
  0.083   bc10   bf7d   be7d   bd9d   bd94   becf   b884   b67d   be78   c859
  0.100   3e00   3f15   3b34   3f71   3d9a   3c2a   3b1c   3b1b   3d83   43ac
  0.117    e54    ea4    ce2    e0a    e16    d58    d8d    cd8    dcb    fff
  0.133    3bf    35f    2f3    323    371    333    328    2eb    32a    3c6
  0.150    205    1b4    14d    193    1c0    17f    14c    175    14e    1a4
  0.167    187    11c    11f    109    154     fd    107    11b     ff    141
```

DDN Slow Disk Drive (3)

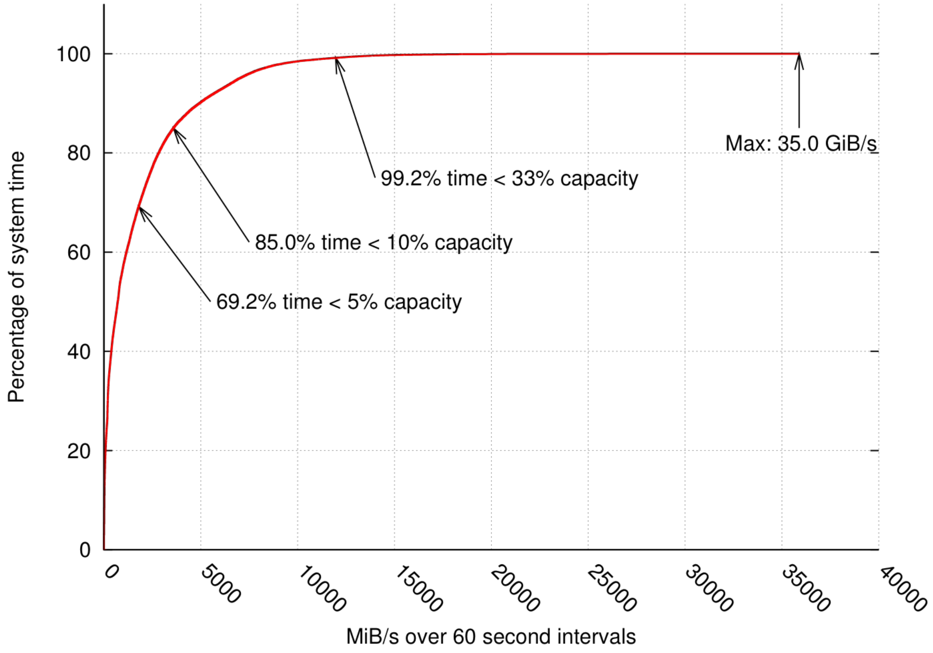


XFP Failures

- The I/O node XFP can degrade before it fails outright with some interesting side effects.
- Transmit side will start to exhibit high packet loss.
- Receive side will continue to operate normally with minimal or zero packet loss.
- GPFS file system client will begin to accumulate locks that are held for extended periods of time because so little data is successful in making it to servers.
- Eventually entire GPFS file system is locked up waiting on this node to release critical resources.
- Performance of a failing node was tested using iperf
  - Tx – 2.0 Mbps
  - Rx – 2.5 Gbps

Free Time

- The I/O node is underpowered with respect to CPU and RAM so care would need to be taken when running any monitoring software on it.
  - Does have a window at job startup / shutdown to run something to collect data
- File servers do have plenty of time available for active monitoring
  - The CDF shows that the system is less then 33% loaded 99% of the time.



Solutions?

- All of the numbers I listed before will be larger for the next system except the number of people who will be working on the system.
- Need some automated system to find when problems are happening and give a root cause of what is causing the problem.
  - List of sense points that provide valuable data
  - Framework to evaluate these sense points to see if they can predicate faults
- All of these problems were debugged/found by running a benchmark across the whole system then reviewing all the statistics that are available.
  - Lengthy process to find issues manually
  - Key take away is that we know this benchmark uses the whole system and distributes the same amount of data to every node.
- How to handle variable work loads which are not benchmarks?
- Magellan
  - http://magellan.alcf.anl.gov/

Future

- Collaboration with Priya Narasimhan and Mike Kasick of CMU
- Work with detecting performance loss and isolating root cause through monitoring of I/O subsystem.
- “Black Box” approach that doesn’t depend on a particular parallel file system.
- Hope to deploy some type of solution for use during bring up of the next ALCF BG/Q machine.
- **Black-Box Problem Diagnosis in Parallel File Systems**
Michael P. Kasick, *Carnegie Mellon University; Jiaqi Tan, DSO National Labs, Singapore; Rajeev Gandhi and Priya Narasimhan, Carnegie Mellon University* (FAST 2010)

Acknowledgements

- Loren Wilson (LCF)
- Phil Carns (MCS)
- Priya Narasimhan (CMU)
- Mike Kasick (CMU)
- Bill Allcock (LCF)
- Rob Ross (MCS)